Title: Extractive Summarization of Bangla Documents by Fine Tuning BERT Models

Abstract:
In recent years document summarization has become an important task in Natural Language Processing. While much research has been done on automated English text summarization, a handful of works have been done on low resource languages such as Bangla. Extractive summarization is the process of extracting important words from a document and then concatenating the important parts to create a shorter representation of a larger document. In this paper we propose to create an extractive summarizer for Bangla texts or documents. We aim to use the Bengali News Summarization dataset built by Bhattacharjee et. al. [1]. For encoding we use a BERT model and then add summarization layers to the output of the BERT model.The summarization layer consists of a sigmoid function for calculating the score of each sentences, a transformer layer in between the sentences to extract important points and finally a Long-Short Term Memory (LSTM) layer for learning summarization specific features.The final layer is a sigmoid function layer for classification and the final predicted score of the sentences is calculated. Based on the score, the top 30% sentences are selected as candidates for the summary paragraph. We experiment using multilingual m-BERT and XLM-RoBERTa and compare the performance of both the models.

Keywords: BERT, Transformer, LSTM, Summarization, Bangla, NLP

Reference:
[1] Bhattacharjee, P., Mallick, A., Islam, M.S., & Jannat, M. (2020). Bengali Abstractive News Summarization(BANS): A Neural Attention Approach. *ArXiv, abs/2012.01747*.
[2] Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. *EMNLP/IJCNLP*.
[3] Liu, Y. (2019). Fine-tune BERT for Extractive Summarization. *ArXiv, abs/1903.10318*.
[4] Sen, O., Fuad, M., Islam, M.N., Rabbi, J., Hasan, M.K., Fime, A.A., Fuad, M.T., Sikder, D., & Iftee, M.A. (2021). Bangla Natural Language Processing: A Comprehensive Review of Classical, Machine Learning, and Deep Learning Based Methods. *ArXiv, abs/2105.14875*.
[5] Haque, M., Pervin, S., Hossain, A., & Begum, Z. (2020). Approaches and Trends of Automatic Bangla Text Summarization: Challenges and Opportunities. *Int. J. Technol. Diffusion, 11*, 1-17.